Claims

I claim:

Sub B1

1. A method of providing load balancing among host servers in a computer network using a load balancing switch and a plurality of site switches, the method comprising:

coupling said load balancing switch between said computer network and an authoritative domain name server and configuring said load balancing switch as a proxy to said authoritative domain name server;

coupling each of said host servers to said computer network through said site switches;

collecting at said load balance switch a first set of performance metrics regarding said network;

whenever said authoritative domain name server provides network addresses in response to a query regarding a domain name, arranging said network addresses as an ordered list in accordance with said performance metrics;

forwarding said ordered list of network addresses as a response to said query to an originator of said query.

2. A method as in Claim 1, further comprising:

collecting a second set of performance metrics regarding said network, said second set of performance metrics reflecting access conditions

to said host servers at each of said site switches;

sending said second set of performance metrics from said site switches to said load balancing server; and

including said second set of performance metrics with said first set of performance metrics.

3. The method of Claim 1, wherein said first set of performance metrics includes a health check sent from said load balancing switch to each of said site switches.

4. The method of Claim 3 wherein, when any of said host servers fails said health check, a network address of said failed host server is provided a lesser position in said ordered list.

5. The method of Claim 2, wherein said collection of said second set of performance metrics includes recording, at each site switch, a number of sessions connected to host servers having network addresses configured on said site switch.

6. The method of Claim 5, wherein when said number of sessions at said site switch exceeds a predetermined percentage of that site switch's maximum capacity, a corresponding one of said network addresses is provided a lesser position in said ordered list.

7. The method of Claim 2, wherein said collecting said second set of performance metrics

includes recording, at each site switch, a round trip time indicative of elapse time for exchanging messages between each site switch and a client machine of said computer network.

8. The method of Claim 7, wherein said round trip time being an actual recorded time period between said site switch receiving a connection request from said client machine and said site switch receiving an acknowledgement of a connection from said client machine.

9. The method of Claim 1, wherein said arranging takes into consideration the geographical location of said originator of said query.

10. The method of Claim 3, wherein said collecting of said first set of performance metrics includes recording a time interval for each site switch between said load balancing switch initiating said health check and said load balancing switch receiving a response from said site switch.

11. The method of Claim 1, wherein said arranging selects a network address of a least recently selected host server for placement at a higher position in said ordered list.

12. The method of Claim 1 further comprising said load balancing switch limiting a valid time for each network address in said ordered list to less than a predetermined value.

13. The method of Claim 1, further comprising, when a connection request is received at a site switch for a connection to one of said host servers, said site switch redirecting said connection request to another one of said host servers.

14. A system for balancing load among host servers in a computer network, comprising:

an authoritative domain name server;

a load balancing switch coupled to said authoritative domain name server, said load balancing switch (a) being configured to be a proxy to said authoritative domain name server; (b) collecting a first set of performance metrics regarding said network; and (c) arranging a list of network addresses from said authoritative domain name server in accordance with said first set of performance metrics; and

a plurality of site switches coupling said host servers to said computer network.

15. A system as in Claim 14, wherein each of said site switches (a) collects a second set of performance metrics regarding said network, said second set of performance metrics reflecting access conditions to host servers at said site switch; and (b) sends said second set of performance metrics to said load balancing server; whereupon said load balancing switch includes said second set of performance metrics with said first set of performance metrics.

16. A system as in Claim 14, wherein said first set of performance metrics includes a health check sent from said load balancing switch to each of said site switches.

17. A system as in Claim 16 wherein, when any of said host servers fails said health check, a network address of said failed host server is provided a lesser position in said ordered list.

18. A system as in Claim 15, wherein said second set of performance metrics includes a number of sessions connected to a network address configured at said site switch.

19. A system as in Claim 18, wherein when said number of sessions exceeds a predetermined percentage of that site switch's maximum capacity, said network address is provided a lesser position in said ordered list.

20. A system as in Claim 15, wherein said second set of performance metrics includes a round trip time indicative of elapse time for exchanging messages between each host server coupled to said site switch and a client machine of said computer network.

21. A system as in Claim 20, wherein said round trip time being an actual recorded time period between said site switch receiving a connection request from said client machine and said site switch receiving an acknowledgement of a connection from said client machine.

22. A system as in Claim 14, wherein said arranging takes into consideration the geographical location of said originator of said query.

23. A system as in Claim 17, wherein said first set of performance metrics includes a time interval for each site switch between said load balancing switch initiating said health check and said load balancing switch receiving a response from said site switch.

24. A system as in Claim 14, wherein said arranging selects a network address of a least recently selected host server for placement at a higher position in said ordered list.

25. A system as in Claim 14, wherein said load balancing switch limits a valid time for each network address in said ordered list to less than a predetermined value.

26. A system as in Claim 14, wherein when a connection request is received at a site switch for a connection to one of said host servers, said site switch redirects said connection request to another one of said host servers.

add #'